DE-83-03

Development of Driver Education Evaluation Tests

- Summary Report -

Ontario

Ministry of
Transportation and
Communications

Transportation
Technology and
Energy Branch

DE-83-03

Development of Driver Education Evaluation Tests

- Summary Report -

Ontario

Ministry of
Transportation and
Communications

Transportation
Technology and
Energy Branch

Development of Driver Education Evaluation Tests

- Summary Report -

Principal Investigators

G.R. Engel
M. Townsend
Engel & Townsend Consultants
Toronto, Ontario

Project Monitor

L.V. Clifford
Research Officer
Human Factors Section, MTC

Prepared for
Safety Co-ordination & Development Office
Transportation Regulation Branch

Published by
The Transportation Technology and Energy Branch
Ontario Ministry of Transportation and Communications
Hon. James W. Snow, Minister
H.F. Gilbert, Deputy Minister

For additional copies, write:
The Editor, Technical Publications
Ontario Ministry of Transportation and Communications
1201 Wilson Avenue
Downsview, Ontario
Canada M3M 1J8

November 1983

ABSTRACT

This summary report gives the general reader an overview of how a Driving Knowledge Test and a Driving Situations Test were developed. Two separate reports "Examiner's Manual - Driving Knowledge Test" DE-83-02, and "Examiner's Manual - Driving Situations Test" DE-83-03 contain all the technical details that are needed to understand how the two tests were constructed; how the tests are to be administered, scored, and interpreted; and how to evaluate the validities and reliabilities of the tests.

The two tests were designed to be used in evaluating the Ontario Ministry of Transportation and Communications' high school driver education program. One test is a test of knowledge about safe driving; the other is a test of an individual's sensitivity to accident risks in different driving situations.

Both tests meet standards of acceptable validity and reliability as defined by modern testing literature, and by industrial usage. The tests should be useful as intermediate instruments for evaluating driver education.

ACKNOWLEDGEMENTS

This project could not have been completed without the generous support and help of many individuals. The following people have our sincerest thanks.

Barry Betzner (Toronto Transit Commission); Barry Bragg (Abt Associates); Ed Blake (Ministry of Transportation and Communications); James Bookbinder (T.T.C.); Victor Bridgette (T.T.C.); Linda Clifford (M.T.C.); R.G. Crothers (Alberta Transport); David Duncan (M.T.C.); Audrey Foden (M.T.C.); Ralph Gallienne (M.T.C.); Bill Johnson (M.T.C.); Brian Jonah (Transport Canada); William Keen (M.T.C.); Al Manly (Thornlea Secondary School); Angus MacFarland (T.T.C./Amalogomated Transit Union); Gordon Nakashima (M.T.C.); Phil Randell (Driver Education Consultants/Donhead Secondary School); Barbara Rowe (York County Board of Education); Roy W. Strickland (T.T.C.); Lakerem Sukhu (Ontario Motor League); Bill Thomson (M.T.C.); Linda Tonelli (M.T.C.); Bill Towne (Donhead Secondary School); Paul Wake (M.T.C.).

We would also like to thank all of the people who volunteered to act as subjects for this project: Toronto Transit Commission Drivers; Driver Education Students at Donhead, King City, Markham District, Sutton District, and Thornlea Secondary Schools; and many members of the general public.

SUMMARY

The purpose of this project was to produce two tests to be used to evaluate the Ontario Ministry of Transportation and Communication's revised High School Driver Education programme. The resulting tests were: The Driving Knowledge Test which measures knowledge of safe driving, and The Driving Situations Test which measures risk-taking tendencies in different driving situations.

The specifications for both tests required; firstly that the tests' contents reflect experts' opinions about what is important to safe driving; and secondly, that the tests would be able to discriminate among drivers acknowledged to represent different levels of safe driving ability. To meet the first requirement, the tests were designed according to content criteria that were systematically derived from expert opinion. To meet the second requirement, the tests were validated, and crossvalidated, on three groups of drivers; full-time professional fleet drivers, driver education students, and a group called nine-point drivers. The nine-point drivers were drivers who had accumulated nine or more demerit points on their driving records as a result of traffic violations. The validities of the tests were measured by the degree to which scores on the tests would discriminate among drivers from these three groups.

The resulting Driving Knowledge Test has two forms, each containing 60 multiple choice items. The items cover 21 knowledge areas defined by the experts as being relevant to evaluating driver education programmes. These areas range from defensive driving to traffic signs and laws. The test can be administered individually or to groups, and it takes about 30 minutes to complete. The maximum score is 60. Average professional drivers score about 50; average students score about 35; and

average nine-point drivers score about 40. About 90% of the professionals score higher than about 90% of students. The validity of the test is about 0.75, for discriminating between professional and student drivers. The coefficient of internal consistency, or reliability, is about 0.85.

Since validity coefficients were calculated for discriminating among and between all three driver groups; for both Forms A and B; and for experimental and crossvalidation administrations of the test, approximate numbers are presented here to give the reader the flavour of the overall results. Complete details can be found in the body of the report.

Reliability coefficients were calculated for both the experimental and crossvalidation administrations of the test and for both Form A and Form B. The reliability coefficient reported here is, once again, a summary statistic. All validity and reliability coefficients are significant at, at least the .01 level.

The Driving Situations Test contains 40 items. Each item gives a scenario describing a driving situation. Each scenario contains three out of five possible dimensions describing; a driver's behaviour, a driver's physiological state, a weather condition, a condition of the vehicle, and a reason for driving. After reading each scenario the examinee rates how much trade-off the driver in the scenario is making between driving and the risk of an accident. The trade-off is rated on a seven point scale.

The score on the test is the arithmetic average of all of the examinee's ratings. The validity of the test is in the range of 0.40 to 0.50 depending on whether the test is discriminating among professional and non-professional drivers, or between students and professional drivers. The reliability of

the test is about 0.96. Once again these values are approximations and complete details can be found in the body of the report. All reliability and validity coefficients for this test are significant at, at least the .01 level.

Scores on the tests are moderately related to each other such that the higher the knowledge score, the more likely is a driver to attribute risk to the various driving situations. Scores on the Driving Situations Test cannot be completely explained by driving knowledge however. That is, there appear to be genuine individual differences among drivers in their willingness to accept risks independent of knowledge about the risks.

Recommendations

The two tests cover both driving knowledge and attitude toward risk-taking in driving situations. Both tests meet standards of acceptable validity and reliability as defined by modern testing literature, and by industrial usage. By these criteria, the tests should be useful as intermediate instruments for evaluating driver education; either for evaluating individual students or for evaluating groups of students.

CONTENTS

LIST OF TABLES

INTRODUCTION

This report describes the development of two tests, the Driving Knowledge Test and the Driving Situations Test. These tests were designed to be used in evaluating the Ontario Ministry of Transportation and Communication's revised High School Driver Education programme. One test is a test of knowledge about safe driving; the other is a test of an individual's sensitivity to accident risks in different driving situations. As evaluation devices, the tests would be used to see whether or not graduates of the new driver education programme demonstrate greater knowledge of safe driving practices, and greater consciousness of driving risks, than graduates of the old programme.

The development of the tests included producing a user manual for each test.[1,2] The user manuals contain all the technical details that would be needed to understand how the tests were constructed; how the tests are to be administered, scored and interpreted; and how to evaluate the validities and reliabilities of the tests. This report will not repeat all the details given in the manuals. Instead it will give the general reader an overview of how the tests were developed.

Development of the tests formed part of a multiphase programme aimed at producing, and evaluating, a new driver education course for Ontario high schools. The first phase of the programme produced an evaluation strategy for evaluating the new course. This included defining driver education effectiveness criteria, reviewing then existing evaluation instruments that could be used to evaluate a new driver education course, and designing specifications for new instruments where existing instruments would not satisfy particular effectiveness criteria. The Driving Knowledge Test and the Driving Situations Test represent two of the new evaluation instruments

identified in the Phase 1 planning study. Details of the Phase 1 Study are described in a report entitled "Revision and Evaluation of Driver Education in Ontario, Phase 1: Development of an Evaluation Plan".(3)

In parallel with the Phase 1 study, a Phase 2 study produced curriculum specifications for the new course. The details of the Phase 2 study are reported in "Revision and Evaluation of Driver Education in Ontario, Phase 2: Preparation of a Curriculum Development Plan".(4) The new driver education course specified by Phase 2 has been developed concurrently with, and independently of, the test development work that will be described here. A brief and readable overview of the work done in both Phases 1 and 2 will be found in "Revision and Evaluation of Driver Education in Ontario: Summary of the First Two Phases of the Study".(5)

The Driving Knowledge Test and the Driving Situations Test have been developed in the context of a single project, using a common test development strategy to meet a common set of objectives. We will begin describing their development by describing these general objectives and the development strategy; then we will describe the construction and validation of each test separately. Following that, we will discuss the roles of the two tests as parts of a test battery. To end the report, we will evaluate the strengths and limitations of the two tests relative to the objectives that they were intended to meet.

DEVELOPMENT OBJECTIVES AND STRATEGY

The Phase 1 study[3] recommended developing a driving knowledge test to assess specific knowledge areas defined by the study. The Phase 1 study also recommended developing the knowledge test by using the same method as had been previously used in developing a Transport Canada driving knowledge test.[6] The Transport Canada test had been designed as a criterion valid test in which the criterion of driving knowledge was how well an individual scored on the test relative to the scores of professional drivers from commercial vehicle fleets.

Although items in the Transport Canada test had been chosen to represent a reasonably wide range of driving knowledge, they had not been chosen with predetermined knowledge areas in mind. Instead, items for the Transport Canada test were chosen primarily for their ability to discriminate between professional drivers and driver education students. This meant that items on the Transport Canada test, while discriminating between professionals and students, might not sample knowledge relevant to safe driving as defined in the Phase 1 study.

The objective of developing a new knowledge test was then, to produce a test that would both cover specific areas of knowledge, and discriminate among different groups of drivers acknowledged to represent different levels of safe driving abilities.

In test construction theory, a test designed to cover predetermined areas of content is said to be a Criterion Referenced test. A Criterion Referenced test usually means that the content of the test has been defined by a panel of

experts. And indeed, it was a panel of experts who defined the driving knowledge areas produced in the Phase 1 study.

A test that can discriminate among individuals representing different levels of ability, or knowledge in this case, is said to be a Norms Referenced test. The Transport Canada test was essentially a Norms Referenced test. Each item in this test was an item for which professional drivers got the correct answer substantially more often than driver education students did.

The main objective in developing the Driving Knowledge Test, and indeed the main challenge, was that of developing a test that would be both a Criterion Referenced test and a Norms Referenced test. The challenge in developing such a test lies in the fact that a test judged by experts to cover all essential areas of driving knowledge can still fail to discriminate among drivers representing different levels of safe driving ability. Conversely, a test that discriminates well among drivers of different ability can still cover areas of knowledge that are not particularly relevant to safe driving.

To give two examples that will help to give some insight into this problem: the most powerful item on the Transport Canada test for discriminating between professional drivers and students was an item that asked when one should renew his or her driver's licence. Virtually all professional drivers got the correct answer to this item, and only about half of the students got the correct answer (the students had never renewed their licences before). This was a good item from a Norms Referenced point of view, but a poor item from a safe driving or Criterion Referenced point of view. An example of just the opposite problem came up in the develop-

ment of the new knowledge test. Experts considered knowledge of the effects of alcohol and drugs to be an important area for evaluating driver education programmes. However, it turned out to be next to impossible to find drug and alcohol items that everyone did not know the answers to. In other words, drivers at all levels seemed to be knowledgeable about drug and alcohol effects, and items on this topic are not very useful for discriminating among drivers at different levels. Items on this topic, while being good Criterion Referenced items, are poor Norms Referenced items.

To put the two examples just given in perspective; both are rather extreme examples. As will be seen later on, the driving knowledge test that emerged from the present project did turn out to meet both the Criterion and the Norms Referenced objectives reasonably well. At the same time however, it is worth noting that in the past, tests of driving knowledge, not to mention other driving abilities, have tended to be developed from a Criterion Referenced approach, and developing a test to meet both objectives was still a new challenge.

Turning to the Driving Situations Test; the Phase 1 study recommended developing a driving attitude test that would measure individuals' expectations of the probabilities of accidents under different driving circumstances. Implicit in the use of such a test as an evaluation device is the idea that a well trained driver should have higher, or at least more realistic expectations of the likelihoods of accidents than poorly trained or untrained drivers do. The specific approach recommended in the Phase 1 study suggested having persons taking the test, literally estimate accident probabilities under different conditions. The rationale for this approach was derived from studies of the effects of public information campaigns on drinking and driving.

The public information studies were ones done on advertising campaigns designed to increase drivers' subjective estimates of the likelihood of being stopped and arrested if they were driving under the influence of alcohol. Research showed that such campaigns increased drivers' subjective estimates of the probabilities of being caught for drinking and driving. Treating driver education as analogous with an advertising campaign, it would be reasonable to expect a driver education course to increase subjective probability estimates not only about being caught for drinking and driving, but also for the outcomes of a variety of other risky driving situations. This was the approach set out as the terms of reference for developing the Driving Situations Test.

Like the specifications for the Driving Knowledge Test the specifications for the Driving Situations Test also included particular areas to be addressed by the test. In this sense, the Driving Situations Test was to be designed as a Criterion Referenced Test. A Criterion Referenced attitude test is however, a potentially trivial test. It is reasonable to expect experts to be able to say with some authority what a driver does or does not need to <u>know</u> about driving, but it is probably not so easy for experts to say how drivers should respond when asked about ephemeral quantities like subjective probabilities. For this reason the Driving Situations Test was developed to be primarily a Norms Referenced test whose strength would be judged on its ability to discriminate among drivers representing different levels of safe driving.

The development of the two tests was designed so that both tests could be treated as parts of a single test battery. This would make it possible to see whether or not the two tests measured different qualities related to the effectiveness of driver education as intended; and to see whether or not

the two tests as a battery would form a more powerful evaluation instrument than either test alone.

Initially, two groups of subjects were identified for the purposes of Norms Referenced validation; professional drivers, and driver education students about to graduate from the present Ontario high school driver education course. The professional drivers were full-time drivers from a commercial fleet. The professionals were defined as the norm for safe driving. Both tests were then to be constructed to discriminate between the professionals and the students. One would then expect graduates of an improved driver education course to get test scores closer to those of professional drivers than graduates from the old course did.

In addition to the student and professional groups, a third group of drivers was included in the validation. These were drivers who had accumulated nine or more demerit points on their driving records as a result of traffic violations. These drivers were tentatively identified as "bad" drivers in contrast to the professionals who represented "good" drivers.

To meet a final development objective, the development plans included a crossvalidation phase in which the tests were given to a new and independent sample of drivers after they had been validated on the original sample of drivers.

The purpose of crossvalidation is to ensure that the validity of a test is not a chance occurance, and that it will remain stable when new groups of people take the test. Crossvalidation, or even validation for that matter, has not been a widely applied procedure in developing driving tests. Nevertheless, there are some notorious examples of tests in other fields that initially appeared to be quite powerful tests, but which turned out to be almost useless when applied to new

groups of test takers. Because of this possibility, crossvalidation was treated as an integral part of developing the present tests.

The objectives set out in developing the Driving Knowledge Test and the Driving Situations Test can be briefly summarized as follows:

- The resulting tests should be both Criterion Referenced (cover areas defined by experts) and Norms Referenced (discriminate among drivers known to represent different levels of driving ability);
- Each test should measure demonstrably different qualities related to safe driving;
- The Norms Referenced validities of the tests should meet the test of independent crossvalidation.

Having set out the background and objectives of the development work, we will now describe the construction and validation of each test in turn; beginning with the Driving Knowledge Test.

CONSTRUCTION AND VALIDATION OF THE DRIVING KNOWLEDGE TEST

The experts in the Phase 1 Study defined the following areas of knowledge as important and relevant to evaluating a driver education course:

- alcohol and drug effects
- driving on curves
- defensive driving
- emergency procedures
- lane changing
- hazard detection
- highway/freeway driving
- intersections
- limited visibility/night driving
- merging
- passing
- pedestrians
- right-of-way
- road conditions
- seat belts
- skid control
- stopping
- surveillance
- traffic signs and laws
- turning
- urban driving

We began the test construction process with the intention of emerging with a test containing two parallel forms containing 50 to 60 items each. This was to be achieved by constructing a provisional test containing about twice this many items; administering the provisional test to a sample of experimental subjects; and then keeping the best items from the provisional test for the final version of the test.

To develop the provisional test we drew on a pool of 1,313 items developed by the Highway Safety Research Institute at the University of Michigan. [7] This pool was culled for items that would fit the knowledge areas defined by the Phase 1

experts. Some items were taken and put into the provisional test without modification; others were modified to suit the purposes of the present test; and where no suitable items could be found from the Michigan pool, new items were written. This process yielded 255 items.

As a formal step in producing a Criterion Referenced test the 255 items were submitted to a panel of experts who were asked to perform a procedure called the Angoff[8] procedure. In the present context, the Angoff procedure consisted of each expert going through the provisional test and for each item, estimating the percentage of minimally qualified driver education graduates who should be able to get the correct answer to the item.

At the same time as they were performing the Angoff procedure, the experts were asked to use a five point scale to rate how relevant each item was to safe driving. They were asked to treat the number "1" on the scale as meaning "not relevant at all to safe driving" and the number "5" on the scale as meaning "very relevant to safe driving". Of the original panel of 10 experts, six completed the task. Three of these were experienced classroom instructors, two were traffic safety research professionals, and one was a senior driver education administrator.

The results of the Angoff procedure were used to evaluate whether or not each item addressed an area that a driver should know, and to evaluate each item's difficulty relative to what a driver could be expected to know. The experts' relevance ratings were used to supplement these evaluations. From the Angoff results and the relevance ratings, some of the provisional items were eliminated, and a few more were modified according to suggestions made by individual experts. The end result of this was a pool of 244 items which we divided into two roughly parallel tests of 122 items each. We will call

these tests, Form A and Form B of the experimental test.

The experimental test was administered to three groups of subjects: Professional drivers, Nine-point drivers, and Driver Education Students. The professional drivers were full-time fleet drivers from an organization whose drivers must meet and maintain high standards of proficiency and safe driving. The nine-point drivers were drivers who had just completed an interview with a Driver Improvement Counsellor at one of the Ontario Ministry of Transportation and Communications Driver Control Centres. These were individuals who had accumulated nine or more demerit points on their driving records as a result of traffic violations.

The students in the sample were from a high school driver education course sponsored jointly by the Ontario Ministry of Education and the Ministry of Transportation and Communications. The course was taught in the school year 1981-82 and was based on the text "Power Under Control". This course is the standard one for Ontario high schools, and it is comparable with high school driver education courses given throughout Canada and the United States.

For purposes of validation, the students took the experimental test on either the last, or last but one class of their course. In addition, the students took one form of the test, either Form A or Form B, on the first day of their course.

The experimental sample consisted of 150 professionals, 150 nine-point drivers, and 150 students. The professionals were drawn at random from an employee roster. The nine-point drivers were asked to volunteer to take the test at the time they completed an interview with a Counsellor. The students came from classes whose instructors volunteered to cooperate

in this project. The professionals and nine-point drivers were paid for writing the tests. The students were paid depending on whether or not local school board policy permitted payment.

Analysis of the Experimental Test Results

The purpose of this analysis was to identify the items that would go into the final version of the test. The analysis was done by a process called item analysis. An item analysis consists of finding each item's validity and reliability. For present purposes, an item would be valid to the extent that professional drivers tended to get the correct answer more often than either students or nine-point drivers did. The reliability of an item amounted to the extent to which getting the correct answer on an item corresponded to an individual getting a high total score on the test. Constructing the final version of the test consisted of retaining those items that had high validities and acceptable levels of reliability.

There is a certain amount of judgement involved in choosing items in this way. The relationship between item validity and item reliability is a complex one. Demanding that all items have high item reliability (that is, high correlations with the total test score) will generally yield items all of which measure a single factor or narrow range of ability. Restricting the range of ability measured by a test typically reduces the test's validity. For example, a test that samples a wide range of driving knowledge is likely to be better at disciminating among groups of drivers (hence have high validity), than a test that only covers a narrow range of driving knowledge. Therefore, items were chosen which have high validity and moderate but not too high reliability.

In addition to choosing items according to the item analysis, the selected items were also checked against the experts' Angoff

and relevance ratings. By and large items selected on the basis of the item analysis were confirmed by the experts' opinions. However, a few items, on alcohol and seat belts for example, were retained even though they were not particularly valid for discriminating among the driver groups. The items were left in because these topics were considered important for evaluating driver education by the experts in the Phase 1 Study.

In doing the item analysis, we also paid attention to the potential problem of differential validity. It is conceivable that an item might, for example, discriminate well between professionals and students but not between professionals and nine-point drivers. Similarly, another item might discriminate well between professionals and nine-point drivers but not between professionals and students. As it turned out, for virtually every item professionals most often got the correct answer, followed by nine-point drivers, followed by students. Indeed, the number of nine-point drivers getting the correct answer to an item consistently fell about mid-way between the numbers of professionals and students getting the correct answer. Consequently, differential validity was not a concern in selecting items.

In a test like this one, the overall validity and reliability of the test is a function of the validities and reliabilities of its individual items. Thus, the process of disgarding items from the experimental test included calculating the overall validity and reliability of different possible versions of a final test. This amounted to a fine-tuning process to maximize the validity and reliability of the test as a whole, and to make the validity and reliability of each test form as nearly equal as possible.

The item analysis yielded a final version of the test consisting of two forms, Form A and Form B, containing 60 items each. Both forms of the final version of the test were then administered to three new groups of drivers; 75 Profess-

ionals, 75 Nine-point drivers, and 75 Students. This administration constituted a crossvalidation to see whether or not the test's validity estimated from the experimental sample would be confirmed by the crossvalidation sample.

Validity of the Knowledge Test

In this section we will describe the major results of the validation and crossvalidation phases of the test development. Readers who want to know these results in more detail should look at the User's Manual for the test. We will begin be describing the results relevant to the Criterion Referenced validity of the test; these are the experts' relevance ratings and the Angoff ratings.

Over 95% of the items on the final version of the test have an average relevance rating of three or greater on the five point scale. There is no item that any single expert rated as not relevant. We also calculated the correlation between each expert's relevance ratings and every other expert's relevance ratings. These correlations varied from around 0 to just over 0.50. Thus while the experts generally agreed that all of the items were relevant to safe driving (scores on the five point scale), they did not show strong agreement with each other on exactly how relevant any one individual item might be; as indicated by the low to moderate inter-item correlations.

Correlations among the experts' Angoff ratings ranged from 0.20 to 0.57 (0.41 to 0.57 if one particular expert's ratings were ignored). These correlations represent moderate but not striking agreement about the level of difficulty that any particular item on the test should represent to a minimally qualified driver education graduate.

We also calculated correlations between the experts' Angoff ratings and the percentages of correct responses actually

obtained by the experimental driver groups. The correlation between the experts' Angoff ratings and drivers' scores on individual items was about 0.45 for Form B and about 0.75 for Form A. From these correlations, the experts were moderately to fairly good at predicting the actual levels of difficulties found for individual items. We have no explanation for why they were better at predicting the levels for Form A than for Form B. However, there was no attempt to construct the two forms to be equivalent in terms of experts' Angoff ratings for individual items.

If the experts' Angoff ratings are averaged over the individual items on a test, one can get what amounts to the experts' opinion about an acceptable passing score on the test. According to this calculation, an acceptable passing score on Form A would be 42.2 (out of 60), and on Form B, it would be 41.0. Despite the experts' variability in rating individual items, the average passing scores are quite similar for the two forms of the test. Comparing this calculation of an acceptable passing score with the results actually obtained by drivers in different experimental groups; 98% of the professionals would have passed the test; about 50% of the nine-point drivers would have passed the test; and only about 20% of the students would have passed the test.

To summarize the Angoff results; we believe it is fair to say that the test can be treated as a respectable Criterion Referenced test. Considerable care was taken to ensure that items on the test covered all of the areas identified as being important in Phase 1. And although the experts varied in their detailed opinions about individual test items, their consensus was that all of the items represent relevant knowledge that a qualified driver should know.

An evaluation of the Norms Referenced validity of the test was made from an analysis of data obtained in the experimental and crossvalidation administrations of the test. We will describe the evidence for validity first in terms of the mean scores obtained by the different driver groups, and then in terms of the validity coefficients describing how well scores on the test discriminated among the different drivers.

Table I shows the means and standard deviations of the scores obtained on each form of the test by the professionals, the nine-point drivers, and the students. The means and standard deviations shown in Table I represent results from the experimental and crossvalidation administrations of the test combined together. Data from the two administrations were combined because each group's mean and standard deviation was almost exactly the same on both administrations.

TABLE I

Means and Standard Deviations

	Form A	Form B
Professional		
Mean	48.53	49.42
S.D.	3.80	4.26
N	225	225
Nine-point		
Mean	42.58	43.50
S.D.	6.85	7.48
N	225	225
Student		
Mean	35.95	34.25
S.D.	6.91	8.31
N	225	225

The data in Table I show that the professional drivers scored on average 12 to 15 points higher than students on either form of the test; and that the nine-point drivers scored about mid-way between the professionals and the students. Looking at the standard deviations, it can be seen that the professionals' scores were grouped fairly closely around their mean score, while the scores of the other two groups were more widely dispersed about their respective means.

Using the standard deviations to calculate how much the scores of the different groups overlapped with one another; it can be shown that about 95% of the professionals scored higher than 95% of the students. Similarly, about 70% of the professionals scored higher than about 70% of the nine-point drivers. In other words, there is hardly any overlap between the scores of the professionals and the scores of the students; and very little overlap between the professionals and the nine-point drivers. From the data in Table I, it is clear that the test makes fairly clear discriminations among the three driver groups, particularly the professionals and the students.

Validity coefficients are another, and indeed the conventional way of expressing a test's Norms Referenced validity. For present purposes, the validity coefficient is the statistical correlation between scores on the test and driver status. Table II (next page) shows the validity coefficients calculated for the experimental, crossvalidation and combined samples for discriminating between professionals and students but not including the nine-point drivers. Values are shown for the experimental and crossvalidation samples separately, and then for the two samples combined.

TABLE II

Professional - Student Validities

	N	Form A	Form B
Experimental	300	.759	.762
Crossvalidation	150	.720	.743
Combined	450	.749	.758

The validity coefficients shown in Table II are, as validity coefficients go, very high indeed. Generally, a good test, a good employment selection test for example, will have a validity of around 0.50. And such tests can be shown to have practical utility with validities as low as 0.20.

The values shown in Table II also show that there was very little shrinkage in the validity of either form of the test in going from the experimental administration of the test to the crossvalidation administration. In the process of keeping or disgarding the items in the experimental version of a test to create a final version of a test, one is bound to keep some items that were chance successes on the experimental administration; and similarly throw out some items that were by chance less successful items. For example, suppose that we had administered the experimental version of the test to just the professionals and the students, and everyone had responded randomly to all of the items. From the laws of chance, we would have found at least some items on which professionals got the "correct" answer more often than students. If we kept those items to form a new test, we would have an apparently valid test. However, administering such a test to a cross-validation sample, also responding randomly, would quickly show that the validity of the edited version of the experimental test was purely a chance event.

Of course this is an extreme example. Nevertheless, the process of constructing a test is always diluted by some degree of chance relative to the genuine validity of the test. Crossvalidation shows the degree to which chance has been operating. If the crossvalidation validity values remain close to the experimental values, one can be confident that the test's validity is stable and relatively undiluted by random effects.

Table III shows validities for discriminating professionals from nine-point drivers, with the students excluded. The values in Table III are substantially smaller than those for discriminating professionals from the students. This reflects the fact that the nine-point drivers' scores were substantially closer to the professionals' scores than those of the students. However, the validity values are still respectable ones.

TABLE III

Professional - Nine-point Validities

	N	Form A	Form B
Experimental	300	.481	.437
Crossvalidation	150	.449	.429
Combined	450	.471	.434

Finally, Table IV (next page) shows the validities obtained for discriminating the professionals from the students and nine-point drivers as a combined group.

TABLE IV

Professionals - Nine-point plus Students Validities

	N	Form A	Form B
Experimental	450	.563	.549
Crossvalidation	225	.530	.529
Combined	675	.555	.544

Taking Tables II to IV as a whole, the validities are consistently substantial, consistently similar for both forms of the test, and remain quite stable from the experimental administration to the crossvalidation administration of the test. In short, the test meets currently accepted standards for Norms Referenced tests.

Reliability of the Knowledge Test

A test's reliability gives an index of its measurement error. Any measuring device is subject to error. A steel tape's reading of the distance between two objects includes the true distance between the objects along with some error due to temperature fluctuations, tape sag, and so on. By the same token, a driving knowledge test score contains an individuals's "true" driving knowledge score, plus some error.

The reliability of a test can be measured in a number of ways, but the most common way is to measure the test's internal consistency. This amounts to seeing how well scores on half of the items in a test correlate with scores on the other half of the test. Table V (next page) shows this split half reliability calculated for each form of the test, and for the experimental, crossvalidation, and experimental plus crossvalidation administrations combined.

TABLE V

Split-half Reliabilities

	N	Form A	Form B
Experimental	450	.852	.888
Crossvalidation	225	.837	.866
Combined	675	.848	.888

Relative to typical test reliability values, the values shown in Table V are somewhat low. Generally, one likes to see reliability values of 0.90 or higher. However, low reliabilities tend to be associated with high validity values. To be highly valid, a knowledge test will generally have to be one that samples a broad range of knowledge. As the range of knowledge that the test samples increases, the greater are the chances that knowing the answers to some items will be unrelated to knowing the answers to certain other items. This in turn leads to lower test reliability. Thus, the validity values achieved with the present test have been at a certain amount of cost in reliability.

The primary importance of a test's reliability is in deciding how much faith to put into the accuracy of an individual score. For example, it can be used to decide whether or not the score obtained by one individual is significantly different from the score obtained by another individual. Since the test was designed for evaluating group performance (whether or not students in one driver education course perform better than students in another course), the reliability of this test is not such an important consideration as it might be in other tests. Therefore, we will not go any further into the details of the test's reliability and its application in interpreting individual scores. These details are given in the test manual however.[1]

This completes the description of the main results for the Driving Knowledge Test. There are still some issues to be discussed, the performance of the nine-point drivers for example. However, it will be more convenient to discuss these as general issues, common with issues arising from the results for the Driving Situations Test. Therefore, we will move on to present the Driving Situations Test results, and return to the general issues when the results of both tests can be discussed together.

CONSTRUCTION AND VALIDATION OF THE DRIVING SITUATIONS TEST

The terms of reference for developing the Driving Situations Test specified a test based on asking an individual for estimates of the probabilities of adverse outcomes of various driving situations. The adverse outcomes were to include accidents, getting traffic tickets, and so on.

Developing a test that would meaningfully measure risk-taking tendencies on this basis presented some serious difficulties. The central difficulty in developing a test focused on probabilities is that risky decisions in driving, or anything else for that matter, are not just a matter of a person's subjective estimates of the probabilities of different possible outcomes, but also a matter of the person's expectations about the gains and losses associated with different outcomes. For example, two people might have exactly the same estimates of the probability of an accident in a particular situation, but come to different decisions about driving in the situation because their subjective perceptions of the overall gains and losses are different. Conversely, the same two persons might have equivalent perceptions of the gains and losses but behave differently because their probability estimates are different. In other words, a test of risk taking has to reflect not just an individual's tendencies to under or over estimate probabilities, but also his or her subjective values and priorities. This was the approach we took to developing the Driving Situations Test.

Taking this approach to the Driving Situations Test also had the advantage that it made the test a reasonably straight forward application of human decision theory.[9] There are in fact a number of theories of human decision making, but all of them conceptualize decision making in much the same way: how people make trade-offs between the probabilities associated

with decision outcomes and the payoff values associated with decision outcomes. For the purposes of the Driving Situations Test, its construction became a matter of creating a test that would measure how different people make these trade-offs in the context of driving. In turn, adopting the most commonly used method of studying human decision making, this meant constructing a test that would present a variety of driving scenarios, each of which would involve an accident risk on the one hand, and on the other hand, certain gains or losses to be made from driving. A person's willingness to accept the accident risk relative to the gain from driving in the scenario would then reflect his or her risk-taking tendency.

To provide a systematic basis for creating scenarios, we began by defining five driving factors. The factors were:

. the physiological state of the driver
. the behaviour of the driver
. the condition of the vehicle
. the weather or road conditions and
. the reason for driving.

To keep each scenario reasonably simple, it was made up using three out of the five possible factors. For example, it might describe the state of the driver, the condition of the vehicle, and the reason for driving. Other scenarios would then contain other combinations of three factors. The Driving Situations Test then, consisted of a series of items, each one containing a scenario for which an examinee would rate the extent to which driving would outweigh or not outweigh the risk of an accident.

The five driving factors were chosen not just to provide inspiration for making up scenarios; they also represent basic factors typically used in analyzing the cause of an accident. As such they served another objective, finding out whether or

not particular factors would have greater or lesser weight in determining a driver's assessment of risk. To achieve this objective the test items were constructed so that different factors and certain of their combinations would appear equally often on the test.

To make the scenarios concrete, each factor was defined in terms of four specific actions or states which were also to embody key attitude criteria defined by the Phase 1 experts. For each factor these were:

The Physiological State of the Driver;

- drugs/alcohol
- fatigue
- visual impairment
- emotional state

The Behaviour of the Driver;

- speeding
- tailgating
- disobeying a sign
- no seat belt

The Condition of the Vehicle;

- brakes
- steering
- tires
- general mechanical problem

The Environment;

- rain
- snow
- fog
- ice

The Reason for Driving;

- no compelling reason
- pleasure
- emergency
- obligation to another person.

Given these definitions, a scenario might then consist of driving while fatigued (physiological state), in rain (environment), in an emergency (reason for driving).

A test that included all combinations of the twenty possible states, taken three at a time, would contain 1140 different scenarios. To systematically reduce the total number of scenarios, and make it possible to analyze the later results in terms of the driving factors; a statistical design called a Balanced Incomplete Blocks [10,11] design was used to make up a reduced number of combinations of factors and states for the scenarios. Details of the combinations that emerged using this design are given in the User's Manual for this test.[2] For present purposes, we need only say that in the resulting 40 item test, each main factor appears a total of six times, each pair of factors appears a total of three times, and each triplet of factors appears once.

To respond to each scenario, the examinee was given a seven point scale representing a trade-off between the risk of an accident and the gain to be made from driving in the situation. The scale was to be used so that the number "1" would represent the judgement that accident risk would greatly outweigh the possible gain to be made from driving; the number "7" would represent the judgement that the gain greatly outweighed the risk; and the number "4" would represent the judgement that accident risk and gain were evenly balanced.

We did not ask any experts to do an equivalent of an Angoff procedure, or any other form of relevance rating of the content of the test. Firstly, it had been relatively easy to construct scenarios to reflect the Phase 1 experts' requirements; so there did not appear to be much to be gained by going to a new set of experts. Secondly, the test was very carefully constructed to balance factors and states among items; so there would be no freedom to add or remove items at the suggestions of a new set of experts. Finally, having constructed the test, we saw that it would be very difficult for an expert to do more than guess as to how drivers should, or would, rate the different

scenarios. Thus, aside from the input of the Phase 1 experts, the Driving Situations Test was designed to be more a Norms Referenced test than a Criterion Referenced Test.

The Driving Situations test was administered to exactly the same drivers who took the Driving Knowledge Test in both the experimental and crossvalidation administrations of the tests. The description of these drivers, and the procedure for administering the test have already been given in connection with the Driving Knowledge Test results, so they do not need to be repeated here.

As will be seen shortly, data from the experimental administration of the test showed that the provisional test had some validity for discriminating among driver groups, and that all of the items on the provisional test contributed about equally well to the test's validity. Therefore, there was no reason to disgard items to produce a final version of the test. Given the highly integrated nature of the test, it was likely that either all of the items would turn out to have some validity, or that none of them would. Since all of the items turned out to have some validity, the original version of the test was crossvalidated without modification.

Validity of the Driving Situations Test

The Driving Situations Test was scored by calculating the average rating that an examinee gave for each item on the test. Thus the minimum score on the test would be 1.0. This would indicate an individual who consistently judged that the accident risk greatly outweighed the gain from driving on every item. A maximum score on the test would be 7.0. This would be the score of an individual who thought that the gain from driving consistently outweighed the risk of an accident. In other words, a low score on the test would represent low risk-taking tendencies,

and a high score would represent high risk-taking tendencies.

Table VI shows the means and standard deviations of the scores obtained by the three driver groups. The data in Table VI are based on combined data for both the experimental and crossvalidation administrations of the test. Data for the two administrations were combined because there were no significant differences between the results for the two administrations.

TABLE VI

Means and Standard Deviations

	N	Mean	S.D.
Professional	225	1.89	.634
Nine-point	225	2.37	.797
Student	225	2.75	.896

It can be seen in Table VI that professionals showed the lowest risk-taking tendency, students the highest risk-taking tendency, with the nine-point drivers falling mid-way between the professionals and the students. The average score of all 675 drivers combined was 2.34, showing that drivers generally considered the scenarios on the test to be on the risky side. This is consistent with the fact that the states used to represent the different driving factors typically reflected adverse driving conditions rather than good driving conditions.

There were also differences in the standard deviations for the three groups. The professionals' scores were grouped closely together; the nine-point drivers' scores were slightly more dispersed; and the students' scores were still more dispersed. The differences among the standard deviations probably reflect, at least in part, the region of the scale used by each group. The professionals, by confining their ratings to the lower end

of the scale, were bound to yield scores with a relatively small standard deviation. In contrast, the students, by using the middle of the scale had room to produce scores with a larger standard deviation.

Using the standard deviations to calculate how much the scores of the different groups overlap with one another; we find that about 75% of the professionals scored lower than about 75% of the students, and about 60% of the professionals scored lower than about 60% of the nine-point drivers. Although there is generally more overlap among the three groups on this test than there was on the Driving Knowledge Test, the separations among the groups are still substantial.

Validity coefficients like the ones calculated for the Driving Knowledge Test, were calculated to see how well the test discriminated between the professionals and the other two groups, and to see whether or not the crossvalidation validities would substantiate those found for the experimental samples. Table VII shows validities calculated for discriminating between professionals and students; and not including the nine-point drivers. These validity values are relevant to the problem of evaluating driver education programmes; the problem this and the knowledge test were designed to address.

TABLE VII

Professional - Student Validities

	N	Validity
Experimental	300	.519
Crossvalidation	150	.407
Combined	450	.485

Table VIII shows the validities for discriminating the professionals from the nine-point drivers.

TABLE VIII

Professional - Nine-point Validities

	N	Validity
Experimental	300	.292
Crossvalidation	150	.351
Combined	450	.313

Finally, Table IX shows validity coefficients for discriminating the professionals from the students and nine-point drivers as a combined group.

TABLE IX

Professional - Nine-point plus Student Validities

	N	Validity
Experimental	450	.386
Crossvalidation	225	.343
Combined	675	.372

The validities shown in Tables VII through IX are all statistically significant at at least the 0.001 level. In addition, there were no significant differences between the validities obtained from the experimental sample and the validities obtained from the crossvalidation sample. For this reason, a validity for the experimental and crossvalidation administrations combined is shown in each Table. Since the version of the test administered to the crossvalidation sample was exactly the same as the one administered to the experimental sample there was no reason to expect any loss in validity due

to discarding items from the experimental test. In fact, in the case of the professionals and the nine-point drivers, the validity went up slightly, though not significantly, on cross-validation.

The validity of the Driving Situations Test for discriminating professionals and students is a value that would be considered moderately good for a test of this sort. The test has only fair validity for discriminating between professionals and nine-point drivers.

Reliability of the Driving Situations Test

Table X shows the split-half reliabilities (like the ones calculated for the knowledge test) for the experimental, cross-validation, and combined experimental and crossvalidation administrations.

TABLE X

Split-half Reliabilities

	N	Reliability
Experimental	450	.960
Crossvalidation	450	.967
Combined	450	.962

Since there were no differences among the reliabilities calculated for the three driver groups separately, we have shown values for the three groups combined. The reliabilities shown in Table X are also essentially equal for both test administrations.

A reliability of 0.96 represents a very high reliability. In addition, the reliabilities of individual items on the test

ranged from 0.50 to 0.70, which also represents quite high values for individual item reliabilities.* All of the data on the test's reliability point to the conclusion that the test is a highly reliable one, and that it measures a single attitudinal quality.

Effects of Driving Factors

We analyzed the data to see whether or not the presence of any particular factor in a scenario, vehicle condition for example, would cause drivers to give higher or lower risk ratings. There were no significant differences among the mean ratings given for different driving factors. In other words, not only were all the driving factors given equal weight as suggested by the reliability data, they also made equal contributions to the total score that a driver got on the test. This merely means that the states used to represent each factor (rain, fog, and the like), were such that the total effect that a particular factor had on the test score was fairly constant. If we had wanted to, we could likely have manipulated the contributions made by different factors by simply defining less risky or more risky states to represent each factor in the scenarios.

This completes the description of the results for the Driving Situations Test. The results show that the test is a moderately valid and highly reliable test for discriminating among drivers on the basis of their relative sensitivities to accident risks. We will now turn to results that describe the relationship between the scores on the Driving Situations Test and the scores on the Driving Knowledge Test.

* Item reliability is the correlation between each item and the total score on a test. (See page 12.)

RELATIONSHIP BETWEEN THE DRIVING KNOWLEDGE TEST AND THE DRIVING SITUATIONS TEST

Since both tests were given to exactly the same drivers, we can show the correlation between scores on one test and scores on the other test. In addition, we can show whether or not using both tests together gives greater discrimination among the driver groups than using either test alone.

The overall correlation between scores on the knowledge test and scores on the driving situations test was -0.380 in the case of Form A of the knowledge test and -0.377 in the case of Form B of the knowledge test. The minus sign in the correlations means that as scores decreased on the Driving Situations Test, scores increased on the Driving Knowledge Test. That is, the more risk a driver assigned to the different driving situations, the higher his or her driving knowledge score; and the less risk the driver assigned to the different driving situations, the lower his or her driving knowledge score.

The correlations between driving situations scores and driving knowledge scores were essentially the same for all three driver groups. The value of the correlation between the two tests is statistically significant and it suggests that there is a moderate relationship between the tendency to perceive risk in various driving situations and knowledge of safe driving. The most sensible interpretation of this result would be that knowledge of safe driving serves to increase risk consciousness.

The partial correlation between scores on the Driving Situations Test and driver status, with knowledge held constant, gives a measure of the test's ability to measure risk-taking independent of driving knowledge. The partial correlation found for discriminating the professionals from the students (combined experimental and crossvalidation data) was 0.32.

This value shows that, while driving knowledge makes a substantial contribution to scores on the Driving Situations Test, the Driving Situations Test scores nevertheless represent what could be called a valid and independent measure of driver risk-taking attitudes.

Since the Driving Knowledge Test and the Driving Situations Test measure independent attributes, to a certain extent at least, scores on both tests should provide better discrimination among the driver groups than scores on either test alone. In the present case however, combining scores on both tests adds very little to the discrimination that can be obtained by using the knowledge scores alone.

The discrimination that can be obtained by using two tests instead of one test is a maximum when both tests have the same validity values, and when the correlation between the two tests is zero. In the present case, the knowledge test has a much higher validity than the risk perception test, and in addition, there is a moderate correlation between scores on the two tests. As a result, using the scores on the Driving Situations Test in addition to scores on the Driving Knowledge Test adds only marginally to the discriminiation that can be obtained by using the Driving Knowledge Test alone.

We hasten to add however, this does not invalidate using the Driving Situations Test to measure risk-taking tendencies. What we have said about using both tests together is relevant only if the sole purpose of using the tests is to discriminate between professionals and students.

TWO ADMINISTRATIONS OF THE TESTS

A sample of students took the tests both at the beginning and at the end of their driver education course; although only end-of-course scores were used for test validation. The purpose of the before and after administrations was to explore each test's sensitivity to changes in knowledge or risk perception. Whether or not the tests would or would not detect such changes was not part of the validation because there was no *a priori* reason to believe that they should.

In the case of the Driving Knowledge Test, half of the students took Form A of the test and the other half took Form B of this test at the beginning of their course. At the end of the course, all of the students took both forms of the test. Overall, the post-course scores were about five percent higher than the pre-course scores. This difference was statistically significant. As information relevant to considering whether or not this difference is a substantial one, we will mention that the students at the beginning of their course had either just taken or were about to take a knowledge test as part of the requirements for obtaining a learners' licence.

The Driving Knowledge Test was administered before and after the course only for the experimental administration of the test. Students in the crossvalidation administration took the test only at the end of the course. Since the post-course scores of the crossvalidation administration were the same as the post-course scores of the experimental administration, it seems reasonable to suggest that the increase in the experimental group's post-course scores had something to do with the course itself rather than just prior experience with the test.

The students wrote the Driving Situations test before and after their course on both the experimental and cross-validation administrations of the test. In the case of the experimental administration, there was a small statistically significant improvement in scores going from the pre-course to the post-course scores. The improvement was not confirmed by the crossvalidation results; scores at the beginning and end of the course were almost exactly the same. In considering these results however, it should be kept in mind that the current course was not specifically designed to reduce risk-taking tendencies in the sense that they are defined and measured by the Driving Situations Test.

DISCUSSION

Both the Driving Knowledge and the Driving Situations Tests should serve adequately for evaluating a new driver education course. In the case of the Driving Knowledge Test, it covers knowledge areas that experts believe to be important for evaluating driver education and relevant to safe driving. The passing score derived from the experts' Angoff ratings would serve as a target for graduates of the new course to aim for. The target is also consistent with the norms for the test in that 98% of the professionals met the experts' passing criterion, while only a few of the drivers in the other groups met it.

From the norms for the Driving Knowledge Test there is almost no overlap between the scores of the students and professionals. Thus, if graduates of a new driver education course showed substantial improvements in knowledge, there is more than adequate separation between the present scores of students and the scores of professionals for these improvements to be apparent.

The scores of students and professionals are also closely grouped about their respective group averages. This means that the test would be quite sensitive to even small improvements in scores of graduates from a new course. For example, if 100 graduates of a new course scored just one score point higher than the present students, this improvement would be statistically significant. Whether or not an improvement of one point would be considered a material improvement is another matter. In any case, there is little danger that the test would fail to detect material improvements for lack of statistical significance.

Compared with the previous Transport Canada test, the new knowledge test has higher validity and slightly higher

reliability. As a matter of interest, average percentage scores of professionals were almost exactly the same on both tests; about 80%. Students on the other hand, scored almost 15% lower on the new test than students scored on the Transport Canada test. This difference should not be interpreted as a substantive difference in knowledge between the two groups of students. Without norms for the same students taking both tests, there is no way of telling whether this difference is due to differences in test construction or due to differences in knowledge.

The Driving Situations Test, while being quite reliable, does not have the degree of validity that its knowledge counterpart does. On the other hand, its validity is comparable with validities typically found for attitude tests of this sort. In addition, things that can be said about the knowledge test's ability to detect and measure improvements in scores for graduates of a new course can also be said in principle for the Driving Situations Test.

While the Driving Situations Test can adequately serve evaluation purposes, its real value may lie in showing that risky decision making in driving can be treated using well established techniques for studying human decision making in other areas. For example, the test represents an application of Conjoint Measurement Theory which is widely used in marketing to study consumer decision making. In a broader context, the Driving Situations Test might bridge the gap between the considerable body of knowledge that has been developed about risky decision making in other areas, and our relatively unsophisticated knowledge about risky decision making in driving.

Gaining a better understanding of why there is such a weak, sometimes non-existant, relationship between accidents and driving abilities may be an area where the approach taken in

the Driving Situations Test would be useful. It is becoming more and more evident that producing knowledgeable and skilled drivers may be a necessary condition for reducing accidents, but it is by no means a sufficient condition. This is the importance of the finding that the Driving Situations Test measures risk-taking independent of driving knowledge. Indeed, it is a finding that confirms any driver's own experience; year to year, day to day, and even minute to minute fluctuations in one's value systems have a significant impact on decisions about whether or not to accept a particular driving risk. Applying the established methods represented in a test like the Driving Situations Test might help to better understand how these fluctuations work, and how drivers could be helped in controlling them.

As mentioned in describing the relationship between scores on the Driving Situations Test and scores on the Driving Knowledge Test; the Driving Situations Test seemed to measure a quality that was independent of knowledge, but the independence was not strong. To find out whether or not this weakness of independence is general would require more study. Both tests covered much the same content; so it is not surprising that responses to both tests were correlated with one another. In addition, the Reason-for-Driving factor was kept relatively constant so that risk ratings would be more a reflection of the driving hazards in the scenarios than of drivers' value systems. If we constructed a test that focused more directly on drivers' perceived payoffs, we might get results that are less dependent on driving knowledge and at the same time more illuminating about drivers' risk-taking.

To remark on the results of administering the tests at both the beginning and end of the driver education course: the results for the knowledge test suggested that the present course produces at least a measurable increase in driving knowledge as defined by the experts. How this increase should be

evaluated, we leave to others to decide.

There was a small improvement in the driving situations scores during the course, but the improvement was not confirmed by the crossvalidation results. The present driver education course was not designed with goals that would lead to expecting these scores to improve, so it should not be surprising that they did not. Accepting risk is a trade-off between probabilities and payoffs. If students are as influenced by payoffs related to social expectations and needs to develop self-esteem, as all the evidence about teenage development suggests, it may by difficult to design a driver education course that attempts to remove these fundamental influences from the driving scene. We do not intend these remarks to suggest doubt about what a new course might achieve. We do suggest that any measurable achievement should be treated as achievement indeed.

As a final item, we will turn to the performance of the nine-point drivers on each of the tests. When the nine-point drivers were suggested as an experimental group it was thought that they might provide a norm representing inadequate drivers; and that their performance on either test might come out worse than that of the students. Instead, they performed substantially better than the students. Indeed, if students from a new driver education course performed as well as the nine-point drivers performed, their performance would represent a substantial improvement over students from the present course.

From reviewing demographic and driving experience data collected along with the test administrations, it appeared that the nine-point drivers might not be so much inadequate as they are overexposed. The nine-point drivers reported above average

annual distances driven; and they included more commercial vehicle drivers than would be found in a sample from the general population. Thus, some proportion of their accumulated demerit points can be attributed to greater exposure. However, after correcting for distances driven, the nine-point drivers still had an accident rate twice that of the professionals. On the other hand, the accident rate of the nine-point drivers was still lower than the accident rates that would be predicted for the students in two or three years following graduation from driver education. Indeed, one of the students in this study had already been involved in a fatal accident before graduation. While we have no way of knowing how the test scores of the nine-point drivers would compare with drivers from the general public, the demographic and driving experience data suggests that the nine-point drivers might not have been too unrepresentative of the general public.

REFERENCES

1) Engel, G.R. & Townsend, M. Examiner's Manual: Driving Knowledge Test. September, 1982.

2) Engel, G.R. & Townsend, M. Examiner's Manual: Driving Situations Test. September, 1982.

3) Bragg, B.W.E. Revision and Evaluation of Driver Education in Ontario: Phase 1: Development of an Evaluation Plan. Ministry of Transportation and Communications: Toronto, Ontario: September, 1980.

4) Robertson, A.S., King, A.J.C., Pratt, D. & Murdoch,P.A. Revision and Evaluation of Driver Education in Ontario: Phase 2: Preparation of a Curriculum Development Plan. Ministry of Transportation and Communications: Toronto, Ontario: September, 1980.

5) Clifford, L.V. & Deslauriers, B.C. Revision and Evaluation of Driver Education in Ontario: Summary of the First Two Phases of the Study. Ontario Ministry of Transportation and Communications: Toronto, Ontario: October, 1980.

6) Engel, G.R., Paskaruk, S. & Green, N. Driver Education Evaluation Tests. Department of Transport. March, 1978.

7) Pollock, W.T. and McDole, T.C. Handbook for Driving Knowledge Testing. Highway Research Institute, The University of Michigan, Ann Arbor, Michigan: August, 1974 Report No. HSRI - 001590-3.

8) Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed) Educational Measurement. Washington, D.C.: American Council on Education, 1971, 514-515.

9) Coombs, C.H., Dawes, R.M. and Tversky, A. Mathematical Psychology: An Elementary Introduction, Prentice-Hall Inc. New Jersey: 1970.

10) Winer, B.J. Statistical Principles in Experimental Design. Second Edition, McGraw Hill, New York: 1973

11) Green, Paul On the Design of Choice Experiments Involving Alternatives. The Journal of Consumer Research, 1, Sept. 1974, pp. 61-68.